5. Slovensko-hrvatski paralelni korpus

Vesna Požgaj Hadži, Marko Tadić

Slovensko-hrvatski paralelni korpus bio je jedan je od projekata s područja humanističkih znanosti u sklopu Sporazuma o bilateralnoj slovensko-hrvatskoj suradnji u području znanosti i tehnologije između Ministarstva znanosti i tehnologije Republike Slovenije i Ministarstva znanosti i tehnologije Republike Hrvatske. »Težinu« projektu *Slovensko-hrvatski paralelni korpus* daje činjenica da je jedan od 28 prihvaćenih projekata (od 70 prijavljenih) koji je odobrio zajednički slovensko-hrvatski odbor krajem lipnja 1999. u trajanju od 2 godine (2000–2001). Riječ je zapravo o drugom dvojezičnom korpusu za slovenski jezik (prvi je *Slovensko-engleski korpus* Elan, v. Erjavec, 1999a, b). Rad na projektu započeo je početkom ak. god. 1999/2000.[22] U poglavlju prikazujemo početak rada na projektu, njegove ciljeve i svrhu te upozoravamo na probleme vezane prije svega uz nedostatak digitalne građe za korpus.

5.1. Cilj i svrha projekta

Slovensko-hrvatski paralelni korpus bio je zamišljen kao ishodište za suvremena kontrastivna istraživanja dvaju genetski srodnih i susjednih jezika, koji su uspostavljanjem novih država doživjeli određene preinake između ostaloga i zbog promjene državnoga statusa. Ciljevi projekta su ovi:

- sastaviti usporedni korpus slovenskih i hrvatskih originala te odgovarajućih prijevoda (obostrani prijevodi),
- korpus sravniti (*align*) na razini rečeničnih prijevodnih ekvivalenata,
- omogućiti pristup korpusu na internetu putem web servisa.

Istraživanja korpusne lingvistike, koja posljednjih desetljeća 20. st. nadmašuju ostale lingvističke discipline, osobito u području leksikografije (jednojezičnih, višejezičnih i paralelnih korpusa, među koje je pripadao i ovaj projekt) čine temelj za čitav niz raznih vrsta lingvističkih proučavanja. Bilo je planirano da rezultati ovoga projekta nađu svoju primjenu u:

- kontrastivnim proučavanjima (strukturnim i tipološkim) slovenskoga i hrvatskoga jezika na svim jezičnim razinama,
- leksikografskim proučavanjima,

22 Kao partneri projekta pojavljuju se Filozofski fakultet Sveučilišta u Ljubljani i Filozofski fakultet Sveučilišta u Zagrebu. Projekt vode V. Požgaj Hadži i M. Tadić; u njemu s hrvatske strane sudjeluju I. Pranjković, V. Muhvić-Dimanovski, B. Bekavac, S. Fulgosi, K. Šojat; na slovenskoj strani na projektu rade V. Gorjanc, A. Skubic i Š. Vintar.

- Zbog zastarjelih slovensko-hrvatskih (srpskohrvatskih) rječnika (Jurančič [3]1986, [2]1989)[23] projekt je trebao poslužiti kao temelj i poticaj za dvojezičnu leksikografiju. Slovensko-hrvatski rječnik i obratno danas postaje neophodnim normativnim priručnikom, čega su svjesni i mnogi nakladnici.
- Slovensko-hrvatski paralelni korpus predstavlja i temelj za različita leksikografska i leksikološka proučavanja, npr. proučavanja »lažnih« prijatelja (rječnik slovensko-hrvatskih homonima), proučavanja terminologije (terminološki rječnici), proučavanja kolokacija itd.

• znanosti o prevođenju,
  - Izrađeni korpus, objavljen na internetu, zamišljen je kao dragocjeno pomagalo studentima za vježbe iz prevođenja (slovenski-hrvatski i obratno) na Filozofskom fakultetu Sveučilišta u Ljubljani i Sveučilišta u Zagrebu, naročito zbog aktualnosti tekstova koje korpus sadrži. Naime, korpus obuhvaća tekstove nastale posljednjih desetak godina 20. stoljeća, koji u oba jezika predstavljaju noviju jezičnu situaciju od one koja je postojala za vrijeme zajedničke države, a za razliku od rječnika nude i kontekst.
  - Izrađeni korpus dobro će doći i prevodiocima u obje države kao dopunski izvor informacija, posebice u posljednje vrijeme, kada se ponovno pojavljuju potrebe za kvalitetnim prijevodima, a kao što smo već rekli, dvojezični su nam rječnici zastarjeli i gotovo neupotrebljivi.

• razvoju jezičnih tehnologija,
  - (automatsko) traženje termina i njihovih prijevodnih ekvivalenata,
  - strojno prevođenje i strojno potpomognuto prevođenje,

• didaktičkim/metodičkim proučavanjima,
  - ususatvljivanju razlika između dvaju srodnih jezika koje su najčešće uzrokom pogrešaka u učenju/poučavanju jezika,
  - klasifikaciji interferencijskih pogrešaka u učenju/poučavanju hrvatskoga i slovenskoga jezika kao stranih i drugih, prije svega mislimo na fakultetsku razinu u obje zemlje te osnovnoškolsku u Sloveniji, u kojoj se od godine 2000. u 7, 8. i 9. razredu osnovne škole uvodi hrvatski kao izborni predmet (v. poglavlje 16),
  - izradi didaktičkih izvora (kontrastivnih udžbenika, priručnika, gramatika itd.) za sve razine i sve stupnjeve učenja obaju jezika.

23 O nedostacima Jurančičevih rječnika v. Požgaj Hadži, 1998.

5.2. Sastavljanje korpusa

5.2.1. Korpusni parametri

Opseg korpusa na početku je definiran na milijun pojavnica (po 500.000 pojavnica za svaki jezik). Vremenski raspon korpusa obuhvaća tekstove nastale od 1990. do 2001, čime se pokušavalo približiti reprezentativnosti suvremene jezične situacije u oba jezika. Korpus ima ovu žanrovsku strukturu (v. tablicu 1): 40 % pripada stručnim tekstovima, 30 % publicističkim i po 15 % beletrističkim i znanstvenim tekstovima.

Tablica 1: *Struktura korpusa*

Tekstovi		Pojavnice
Beletristički tekstovi	15 %	75.000 pojavnica
Publicistički tekstovi	30 %	150.000 pojavnica
Stručni tekstovi	40 %	200.000 pojavnica
Znanstveni tekstovi	15 %	75.000 pojavnica

U početnim fazama prikupljanja tekstova za korpus preuzimali su se tekstovi isključivo u digitalnom zapisu. Ubrzo smo naišli na »tehničku prepreku« i uočili da takvih tekstova i takve žanrovske strukture nema dovoljno. Kako potpora projektu nije predviđala troškove za digitalizaciju tekstova, morali smo se ograničiti isključivo na tekstove koji su već postojali u digitalnome zapisu. Međutim, dostupnih tekstova u digitalnome zapisu – od kojih je jedan od paralelnih tekstova izvornik, a drugi njegov prijevod – jednostavno tada nije bilo dovoljno. To je dovelo do nemogućnosti ispunjenja planiranoga opsega korpusa i planirane žanrovske strukture. Također, ograničavajući čimbenik bio je i zahtjev da su tekstovi nastali nakon 1990. Tako postavljeni korpusni parametri pokazali su se prekonzervativnima i onemogućili su sastavljanje korpusa u željenome opsegu. Sakupljena građa do kraja 2000. godine obuhvaćala je međudržavne ugovore, beletristiku (romane i priče),[24] tehničku dokumentaciju i priručnike (promet, kulinarstvo, farmaceutika, elektrotehnika i sl.) te turističke brošure. Njezin je opseg oko 150.000 pojavnica u svakome jeziku.

24 Dramski i poetski tekstovi ne uzimaju se u korpus zbog svojih specifičnosti.

5.2.2. Obrada

Nad dijelom sakupljene građe probno je obavljena cijela obrada koja se sastojala od konverzije teksta u jedinstven XML zapis, priređivanja XML dokumenata za sravnjivanje i konačno samog postupka sravnjivanja. Konverzija teksta iz polaznih oblika zapisa (najčešće MS Word 97) obavljena je programom 2XML[25], koji je razvijen u Zavodu za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu u sklopu rada na *Hrvatskome nacionalnom korpusu*. Smjer je konverzije najčešće bio iz RTF u XML.

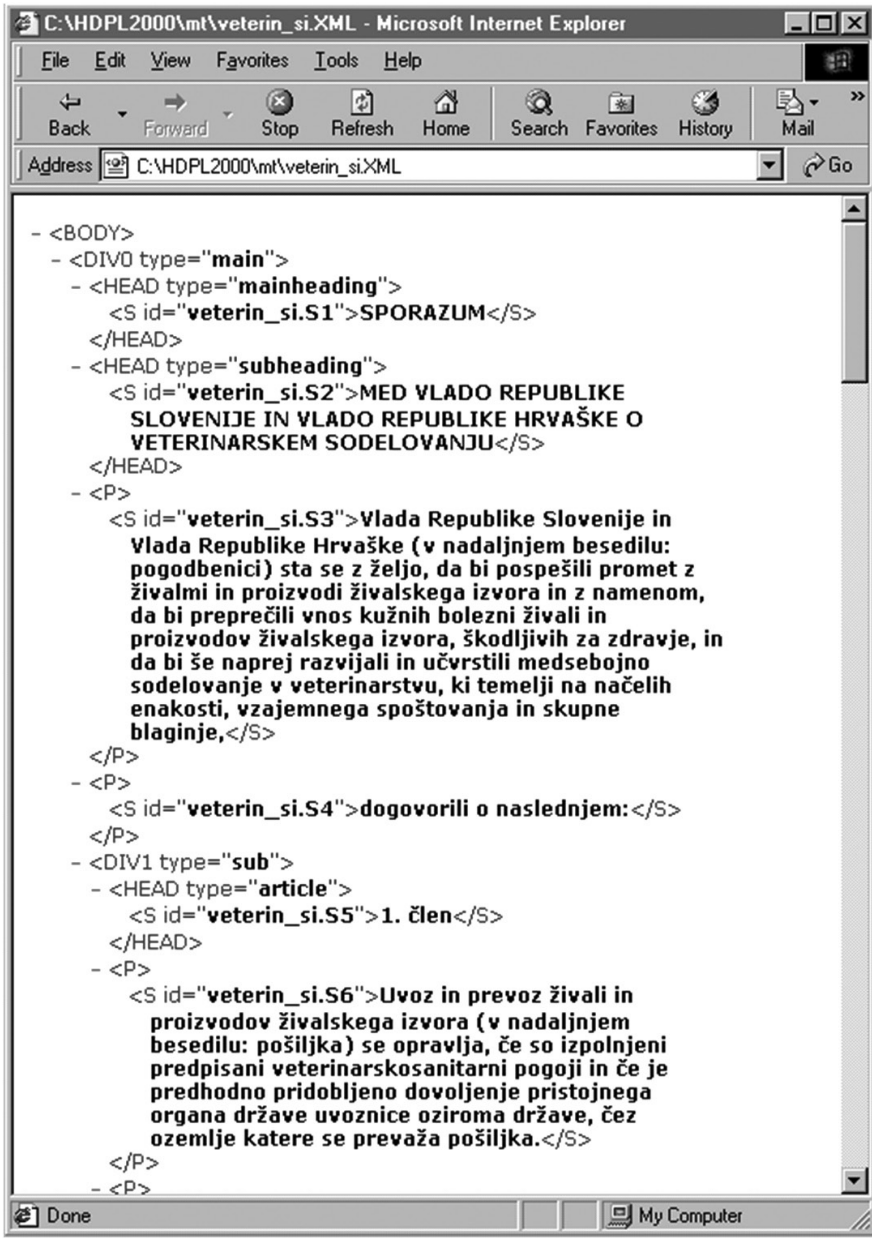
Nakon konverzije u XML obrada dokumenata sastojala se od segmentacije na rečenice <s> uz dodavanje ID oznaka rečenicama te provjere istovjetnosti broja odlomaka <p> kao pripreme za sravnjivanje.

5.2.3. Sravnjivanje

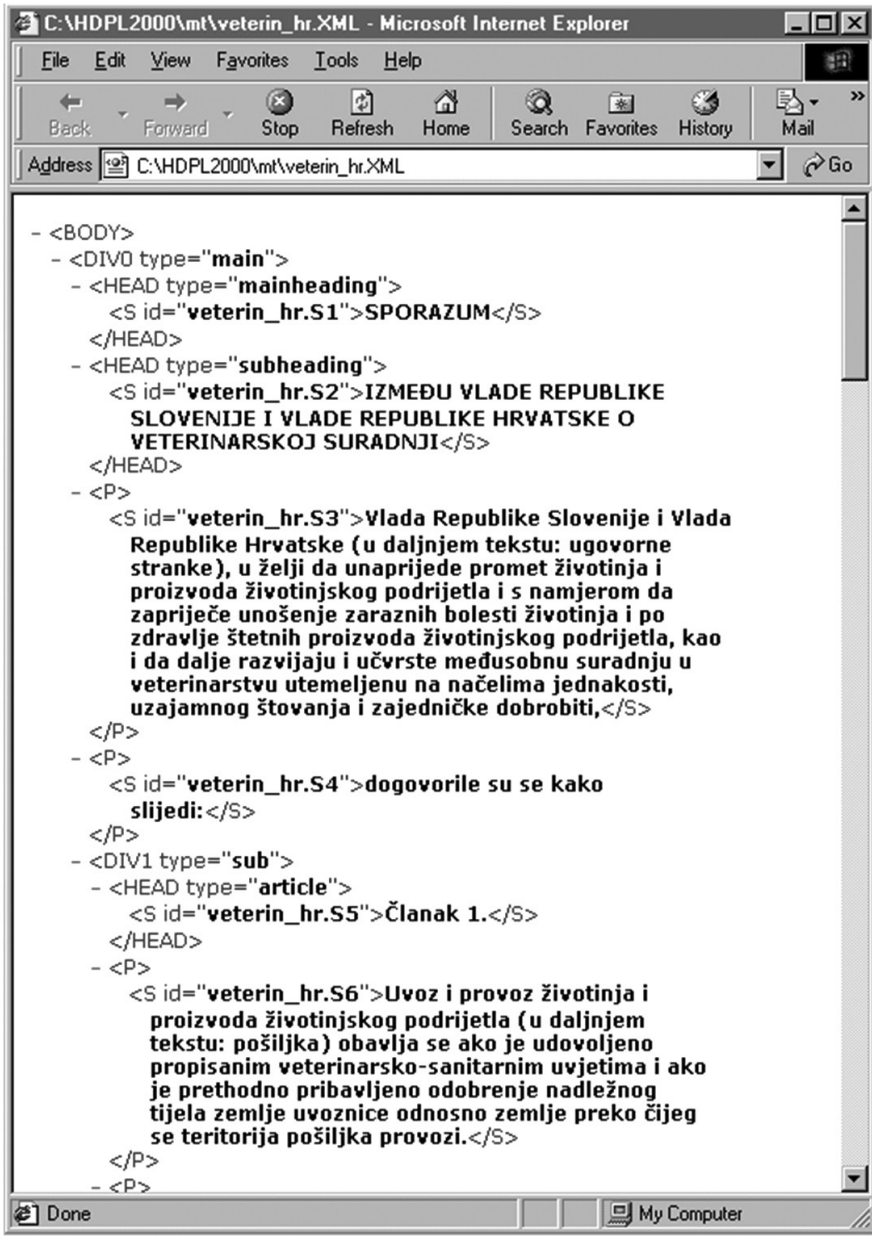
Za sravnjivanje je korištena DOS inačica programa Vanilla Aligner (Danielsson i Ridings 1997) koja se pokazala sasvim dostatnom za potrebe manjega paralelnog korpusa kao što je slovensko-hrvatski. Kako sam program i nema suviše fleksibilan ulazni zapis, XML dokumenti morali su mu se prilagoditi. U tom su procesu međutim zadržane XML oznake te su i one ušle u postupak sravnjivanja. Time je program dobio više redundantnih podataka i mogao je obaviti sravnjivanje na kvalitetniji način.

Probni uzorak čini osam međudržavnih ugovora između Republike Slovenije i Republike Hrvatske. Sami XML dokumenti prikazani su na slikama 1 i 2.

25 Sam program i njegovo funkcioniranje dijelom se prikazuje u Tadić (2000).



Slika 1: *XML dokumenti u probnome uzorku: slovenski ugovori*



Slika 2: *XML dokumenti u probnome uzorku: hrvatski ugovori*

5.3. Rezultati probnoga uzorka

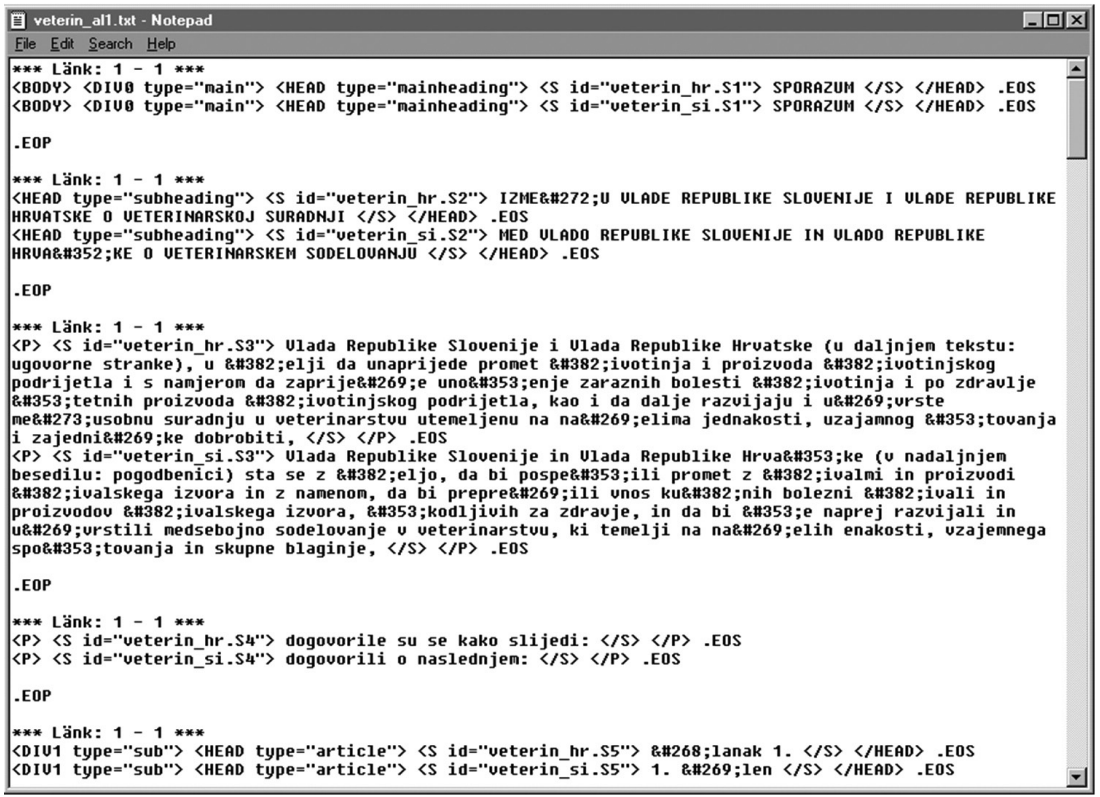
Preliminarna statistika probnoga uzorka osam međudržavnih ugovora između Republike Slovenije i Republike Hrvatske po XML elementima prikazana je u tablici 4:

Tablica 2: *Broj <p>, <s> i <w> elemenata u probnome uzorku*

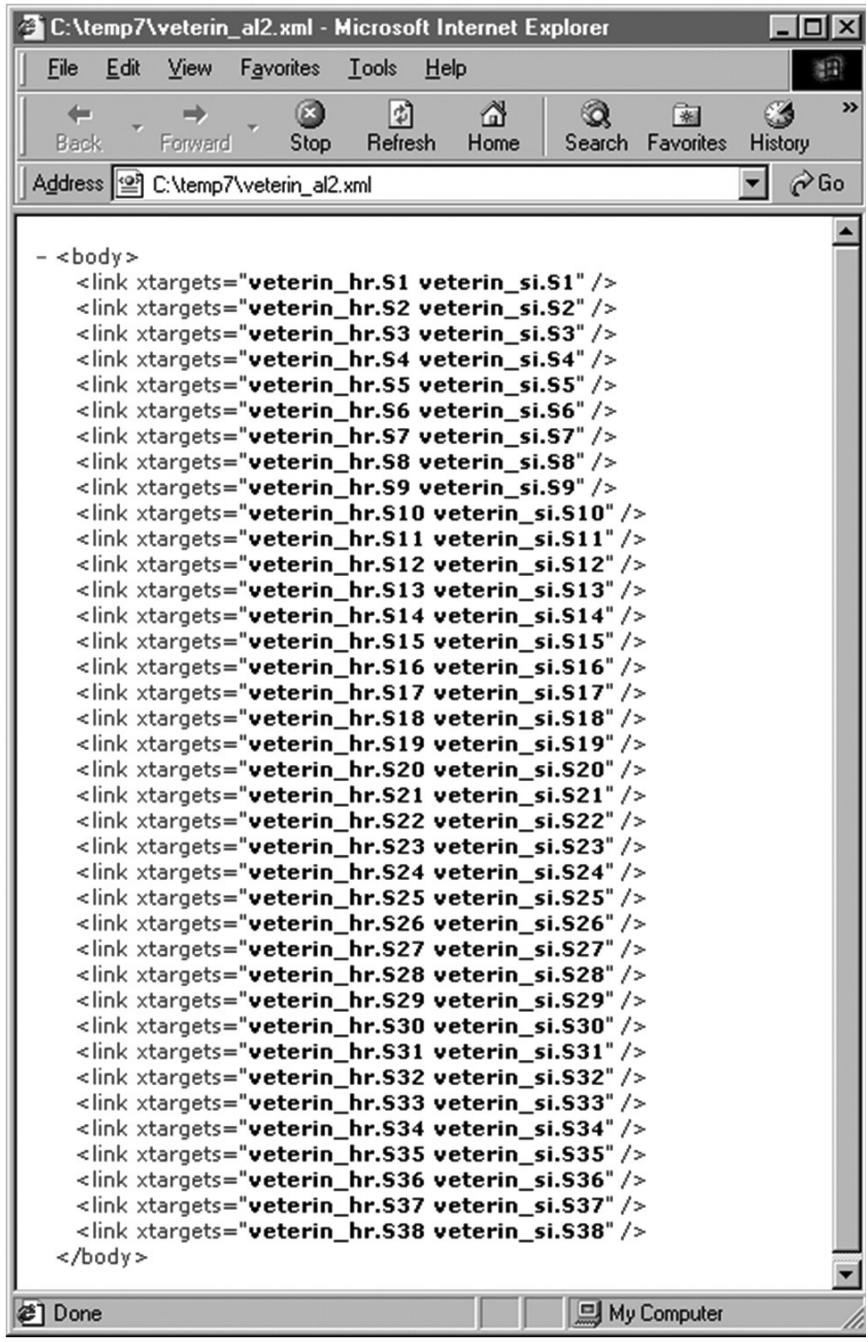
	Hrvatski	Slovenski
Odlomaka <p>	522	522
Rečenica <s>	891	891
Pojavnica <w>	13.549	13.307

Valja uočiti identičnost broja odlomaka i rečenica koja se pojavljuje zbog naravi pravnih tekstova. Tekstovi na oba jezika objavljeni su kao originali i usklađeni već u izvorniku. Stoga su sva strojna sravnjivanja u tom probnom uzorku oblika 1–1. Gotovo identično ponašanje može se naći u drugim paralelnim korpusima pravnih tekstova (v. Gamper, 2000). Dapače, kad se god pri pregledu sravnjivanja naišlo na 2–1 ili 1–2 sravnjivanje, to je redovito bio znak pogreške u ulaznim podacima tj. pogreške u modulu za segmentaciju rečenica.

Rezultat sranjivanja Vanillom prikazan je na slici 3, a u sažetijem zapisu na slici 4.



Slika 3: *Rezultat sravnjivanja Vanillom*



Slika 4: *Rezultat sravnjivanja Vanillom u sažetijem zapisu*

U poglavlju je prikazan početak rada na projektu *Slovensko-hrvatski paralelni korpus* koji je odobrilo i financijski podržalo i slovensko i hrvatsko Ministarstvo znanosti i tehnologije te je time omogućena suradnja stručnjaka dvaju istrovrsnih fakulteta. U prvoj godini rada na projektu (2000) sakupljeni tekstovi, različitih funkcionalnih stilova i različitih područja, zapravo predstavljaju kompromis između očekivane uporabnosti korpusa s jedne strane i same dostupnosti tekstova u digitalnom zapisu s druge strane. Naime, prvih desetak godina nakon osamostaljenja Republike Slovenije i Republike Hrvatske, bilo je uočljivo nepostojanje prijevoda sa slovenskoga na hrvatski i hrvatskoga na slovenski jezik (osobito u digitalnom obliku), što je reazultiralo značajnim teškoćama oko prikupljanja građe za korpus. Kad bude zaokružen u planiranome opsegu i kad postane dostupan putem interneta, *Slovensko-hrvatski paralelni korpus* omogućit će niz različitih jezikoslovnih istraživanja: kontrastivnih, leksikografskih, didaktičko/metodičkih, istraživanja vezanih uz znanost o prevođenju te razvitak jezičnih tehnologija za oba jezika. Potencijalni su korisnici korpusa, osim istraživača i studenata slovenistike i kroatistike, također i prevoditelji kojima će korpus biti dostupan putem www-a kao dopunski izvor informacija, osobito u situaciji kada dvojezičnih hrvatsko-slovenskih (i obratno) rječnika zapravo nema.[26]

26 Korpus, nažalost, zbog teškoća koje smo predvidjeli već u prvoj godini rada na projektu (nedostatak tekstova u digitalnom obliku) nije završen i tako nedostupan korisnicima.